position, or in any other desired manner. Normalization can be desirable to increase the degree to which mutations at a given position are sampled in generating the library.

The phrase "constraint vector" refers to a constraint put on or "applied to" the probability matrix to determine whether and the degree to which mutations at a given position in the matrix are to be included in the library. It too is typically an algorithm that determines whether a particular mutation will result in a functional protein. Variables that can be used to determine the constraint vector are also described below.

II. PROBABILITY MATRIX

A probability matrix is generated to provide an estimate that a given residue will provide a desired activity in a biological polymer of interest. The biological polymer can be a polynucleotide having its own activity of interest, or can encode a protein having an activity of interest. Biological polymers can include polynucleotides exhibiting catalytic activity, for example ribozymes, polynucleotides exhibiting binding activity, for example aptamers, polynucleotides exhibiting promoter activity, or polynucleotides exhibiting any other desired activity, alone or in combination with any other molecule.

The matrix comprises rows representing a given position in the biological polymer of interest, and columns for a plurality of different residues which can be incorporated into the reference sequence. The matrix entries give an estimate for the probability that incorporation of the residue in that column at the position in that row will produce a polymer having the desired activity.

A probability matrix can be generated for at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90 or 100 positions in the reference sequence up to the entire sequence, and can include contiguous residues or noncontiguous residues or mixtures thereof. The matrix can include at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45 or 50 different residues. Naturally occurring residues can be included in the matrix, as well as unnatural residues for synthetic methods, and combinations thereof.

A profile can be created from the matrix based on probability scores and weighting factors. The probability matrix for a protein is preferably an n x 20 matrix that calculates the probability for any point mutation of the target gene that the mutation will result in a protein having the desired function.

In one aspect, a probability matrix is calculated for a given protein library to be produced. To do this, numerical values are assigned to each amino acid that can be substituted into the sequence. One of skill will realize these numbers are arbitrary in that they are relative to each other only for the particular library being produced. It can be useful in some instances to assign the wild type residue at a given position a value of 1, although the wild type residue can be assigned any value. From this initial value, the values of each of the 20 encoded naturally occurring amino acids at each position can be assigned.

In some instances, it can be useful to assume, initially, that the wild type residue is a useful residue and results in a functional molecule. Thus, the value of most other residues should be less than that given to the wild type, therefore in the present example, less than "1". Furthermore, in assigning values, residues that exhibit a low degree of conservation in homologs can be given large values in the probability matrix. Also, because areas of a protein which allow an insertion should be more tolerant to substitution, higher probabilities can be given to nonnative residues at positions which are close to insertions or deletions in homologs.

An example of a ranking of amino acid for valuation in this invention can be found in Gribskov, *Proc Nat'l Acad* Sci *USA* **84**:4355 (1987). The degree of conservation for each position can be used to scale the values according to Gribskov.

Other information can be used to generate a probability matrix. For example, structural information has been found to be useful. As is well known, Hidden Markov models calculate the probability of going from one residue to the next based on sequence alignments. These models also include probabilities for gaps and insertions. See, Krogh, "An introduction to Hidden Markov models for biological sequences," in COMPUTATIONAL METHODS IN MOLECULAR BIOLOGY, Salzberg, et al., eds, Elsevier, Amsterdam.

Other structural information found to be useful is the three dimensional structure of the protein. See for example, Dahiyat & Mayo, *Protein Sci.* 5:895 (1996). This can be determined crystallographically or from molecular modeling techniques. Energy minimization methods can also be employed.

A variety of different substitution matrices can be used as input for the calculation of a probability matrix. The choice of substitution matrix will impact the probability and ultimately the mutagenesis scheme. Thus, if mutations based on sequence alignment are desired, a sequence alignment substitution matrix should be chosen. Alternatively, if mutations that depend on general mutability are desired, a substitution matrix reflecting this need should be chosen.

Substitution matrices can be calculated based on the environment of a residue, *e.g.,* inside or accessible, in α-helix or in β sheet. See, Overington, et al., *Protein Sci* 1:216 (1992). Methods to determine solvent accessible residues are known in the art. *See*, for example, Hubbard, *Protein Eng* 1:159 (1987).

More complex substitution matrices which consider secondary structure, solvent accessibility, and the residue chemistry are also suitable for use in probability matrices. *See*, for example, Bowie & Eisenberg, *Nature* **356**:83 (1992).

One of skill will realize that a probability matrix can require quite complex mathematical calculations and therefore an algorithm that determines the matrix can be desired or even required. The development of such an algorithm is within the skill in the art following the teachings herein. Similarly, because of the complex calculations necessary to carry out the algorithm, it can be desirable to generate a computer program and employ it on a computer to calculate the probability matrix. Again, this is within the skill in the art.

III. CONSTRAINT VECTORS

The constraint vector preferably should reflect the likelihood that a specific mutation at each amino acid position of a protein will improve or affect the desired function of that protein. One example of a constraint vector is a correlation matrix. The constraint vector can also include knowledge-based component(s), such as prior knowledge of effects of single